Universidad
Carlos III de Madrid

Módulo I
Fundamentos Recuperación en Internet

OpenCourseWare

Recuperación y Acceso a la Información

Contenidos

- Buscadores Web
- Tipología
- Internet invisible
- Otros tipos de buscadores
- Tendencias en Internet

¿Qué es un Buscador?

Un buscador es un software que busca en una base de datos o repositorio documental, conforme a algunos criterios y prioridades específicas.

Objetivos

- Indizar la red constantemente para permitir la consulta de sus recursos
- Encontrar los documentos que contengan las palabras clave introducidas por el usuario

Características Web

- Volumen
- Obsolescencia
- Actualización
- Tipología
- Tráfico

Tipos

- Directorios
- Motores de búsqueda
- Meta-buscadores

Contenidos

- Buscadores Web
- Tipología
- Internet invisible
- Otros tipos de buscadores
- Tendencias en Internet

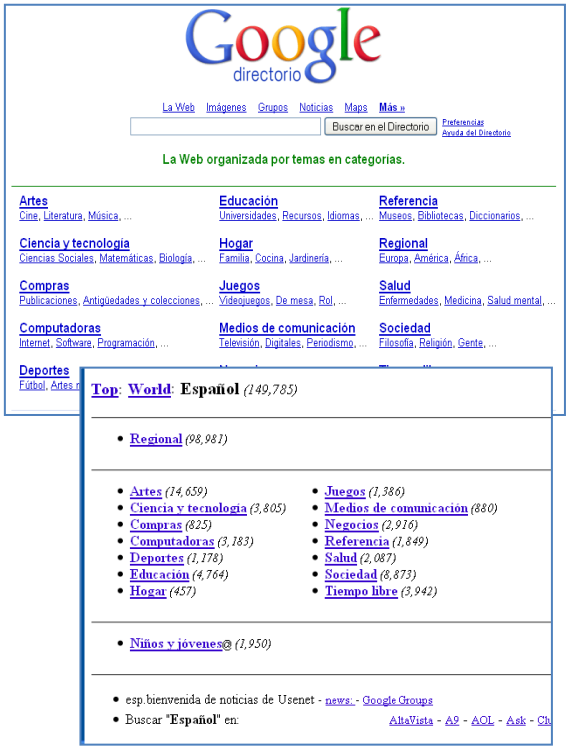
Directorios

¿Qué son?

– Páginas con clasificación intelectual por temas

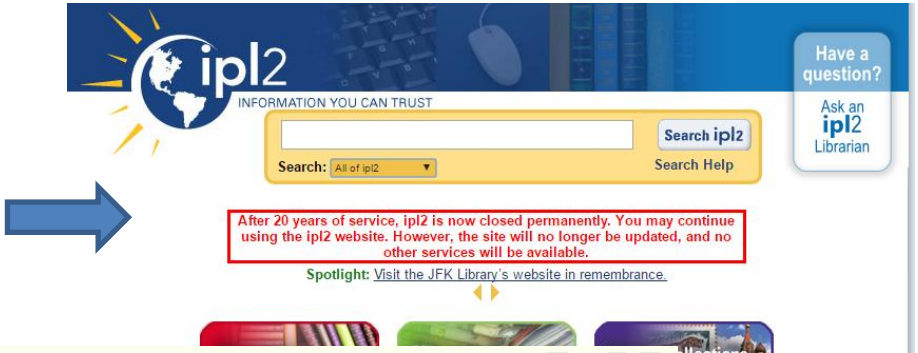
- Características
– Actualización
– Exhaustividad
– Relevancia
– Calidad

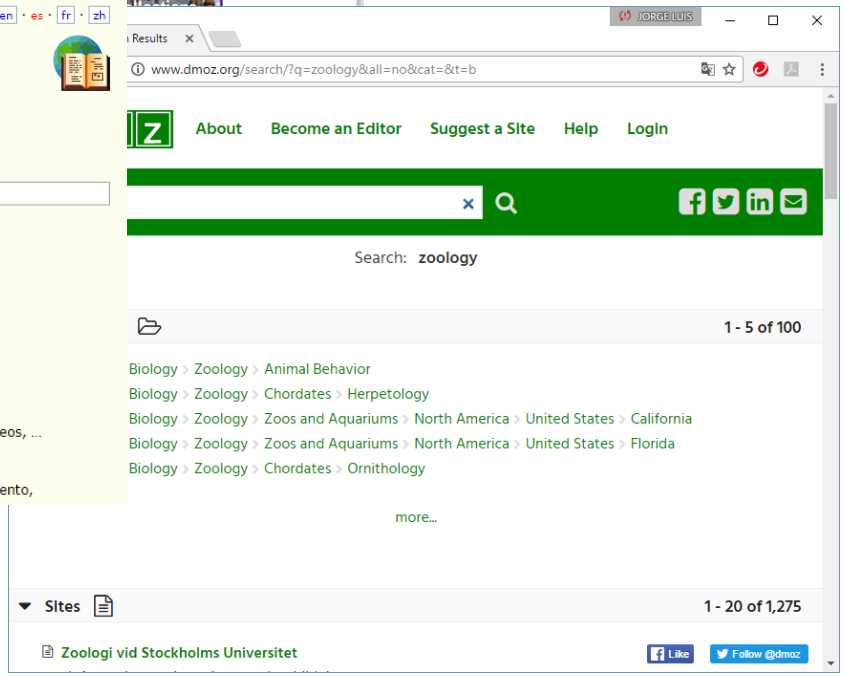
– Ejemplos: DMOZ (1998-2017), Yahoo (1994-2014), Ipl2 (1995-2013), Categorización Wikipedia…



Directorio Yahoo!

Directorios





Motores de búsqueda

- ¿Qué son?

– Aplicación de recolección y búsqueda

- Características

– Exhaustividad

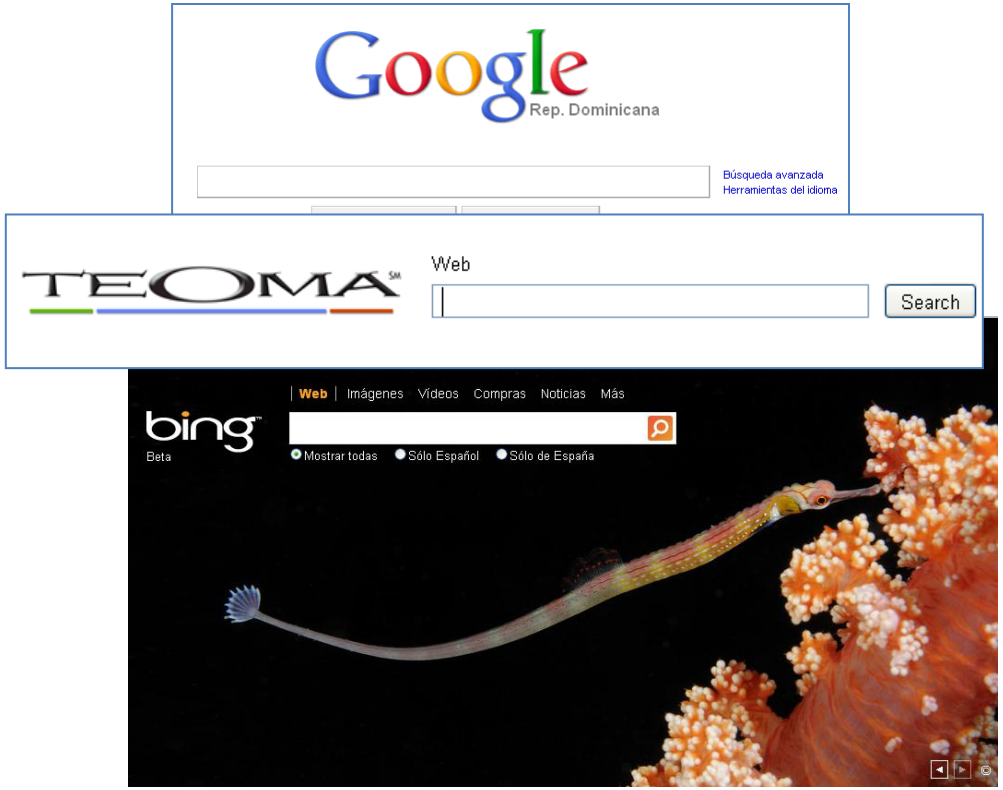
– Actualización

– Manipulación

– Problemas léxicos

– Seguridad

- Ejemplos: Google



Sistema de Recuperación

Recopilación
Documentos

Robots.txt

(texto)

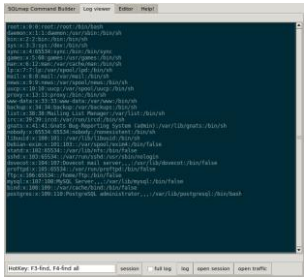
Indización

Bases de datos

Algoritmos de
Recuperación y
Posicionamiento

Reformulación

Consulta
Usuario



Interfaz de
Consulta y SERP

SERP (Search Engine Result Page)

Estadísticas de uso motores

Search Engine	Desktop	Smart phone
Google	77.82%	93.76%
Bing	7.96%	1.06%
Yahoo! Search	6.39%	4.01%
Baidu	6.37%	0.37%
Ask	0.15%	0.04%
AOL	0.08%	0.01%

Fuente https://www.netmarketshare.com/, Dic. 2016

Hay diferencias regionales:

Baidu (China, 55% y Qihoo 360, 28%),
Yandex (Rusia, 58%),
Corea del Sur (Naver, 77%),
Japon (Yahoo Japan, 40%, tras Google)

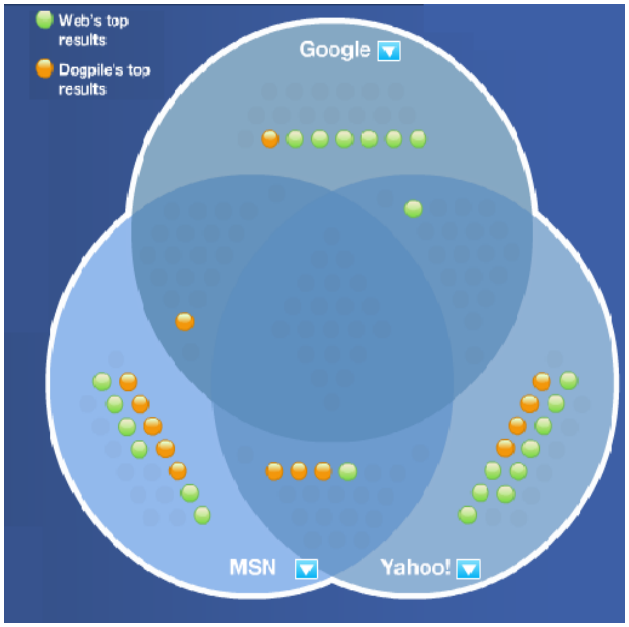
Fuente http://returnonnow.com/ 2015

Metabuscadores

- ¿Qué son?
  - Aplicación que agrega bases de datos, motores y/o directorios
- Características
  - **Algoritmo**
  - **Políticas de uso**

- Diferencia entre:
  1. Mashups
  2. Multibuscadores, agentes de búsqueda, metabuscadores

Ejemplos: Rastreator, Rumbo, Kayak, …



Información almacenada en Google

- Site
- Allinurl
- Allintitle
- Link (no fiable)
- Info
- |
- “”
- -
- 111..333
- Define
- Cache
- Press
- Location
- Index of
- Filetype
- Related

Contenidos

- Buscadores Web
- Tipología
- Internet invisible
- Otros tipos de buscadores
- Tendencias en Internet

Problemas de los buscadores: Internet Invisible

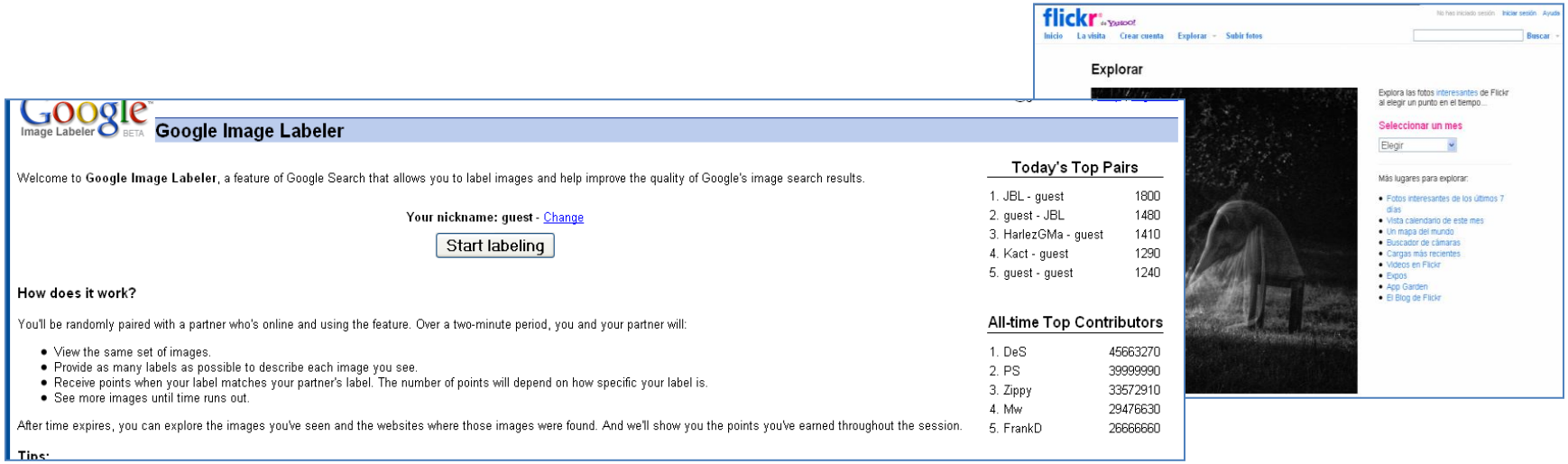
- Carencias de motores de búsqueda debido a:
  - Formatos no soportados
  - Necesidad de validación
  - Formularios de acceso
  - Sitos excluidos expresamente
  - Páginas creadas automáticamente
  - No textual
- No confundir con la DeepWeb
- Aproximadamente tiene un 50% más de tráfico que el visible
- Buscadores especializados: p.e. Easy searcher, OPAC, …

Contenidos

- Buscadores Web
- Tipología
- Internet invisible
- Otros tipos de buscadores
- Tendencias en Internet

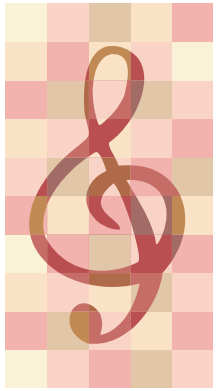
Recuperación de imágenes

- Texto asociado
- Asignación intelectual
- Reconocimiento
- Ejemplos: Google images, Flickr



Recuperación de música

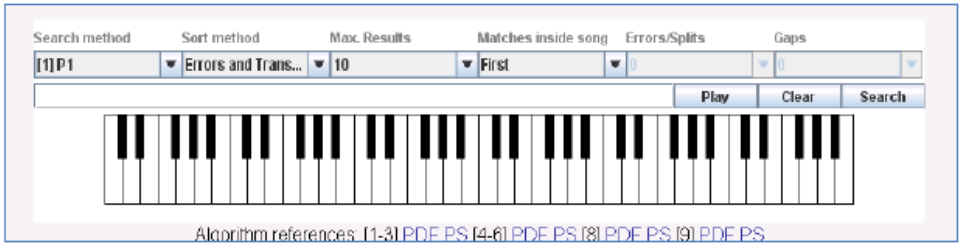
- Similar a texto en origen
- Uso de texto asociado
- Dos tipos: Audio (wav, mp3…) o Simbólica (midi, musicXML…)
- Problemática
  - Especificar consultas
  - Mostrar resultados
  - Derechos



R.Typke et.al., "A Survey of Music Information Retrieval Systems", ISMIR, 2005

Recuperación de música (II)

- Tiene muchas aplicaciones:
  - Identificación
  - Detección de plagio
  - Recomendación
  - Generación de pentagramas
- Hay sistemas a escala industrial, pero la mayoría sigue siendo experimental sobre colecciones pequeñas
  - Shazam
  - C-Brahms
  - Liveplasma



Contenidos

- Buscadores Web
- Tipología
- Internet invisible
- Otros tipos de buscadores
- Tendencias en Internet

Sistemas pregunta respuesta

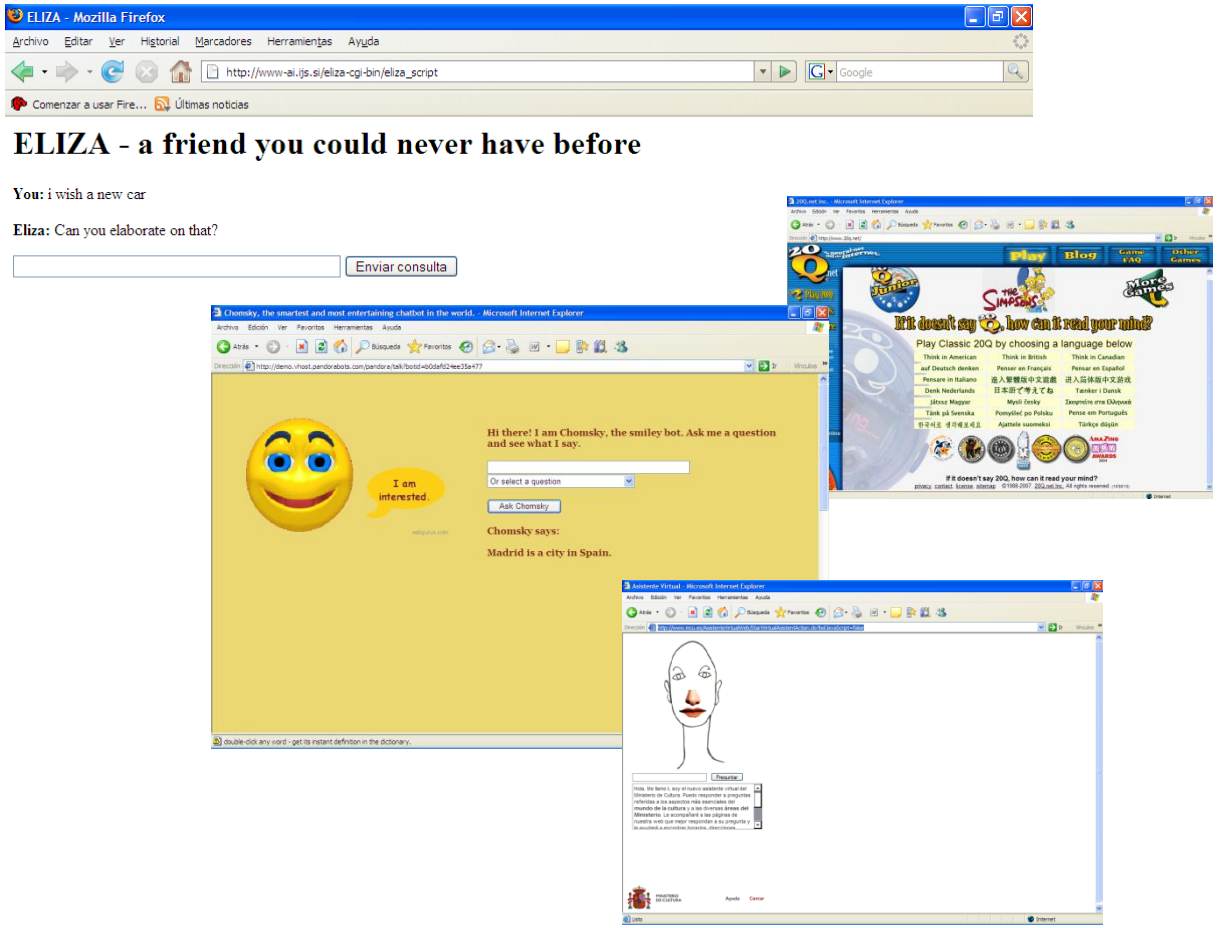
Un sistema pregunta-respuesta devuelve una respuesta concreta a una pregunta y no un conjunto de documentos que puedan contener la respuesta, un subtipo son los asistentes virtuales

- Buscadores con interpretación del lenguaje natural:
  - Asistentes virtuales
  - Answers.com http://www.answers.com/
- Buscadores con interpretación semántica
  - Start http://start.csail.mit.edu/
  - WolframAlpha http://www.wolframalpha.com/

Asistentes Virtuales

Con:

- PLN
- Organización léxica
- Patrones
- Estadísticas



- Ella
- Chomsky
- IKEA (Anna)
- Siri (iOS)
- Renfe (Irene)

¿Qué es la web semántica?

Tim Berners-Lee

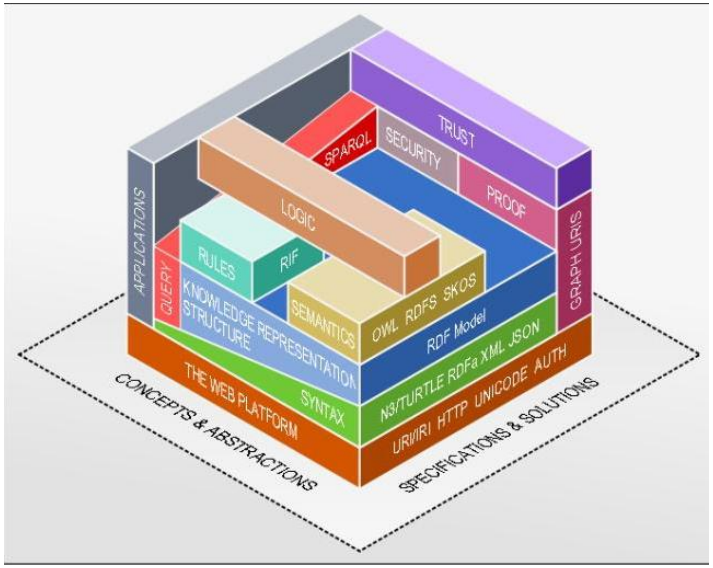
Una Web Semántica es una red de datos que pueden ser procesados directa o indirectamente por máquinas.

Es una web extendida que permitirá a humanos y máquinas trabajar en cooperación mutua.

Guia breve de la Web Semántica http://www.w3c.es/Divulgacion/Guiasbreves/WebSemantica

¿Para qué sirve la Web Semántica?

Procesar contenido, razonar, combinarlo y realizar deducciones lógicas para resolver problemas cotidianos de forma automática

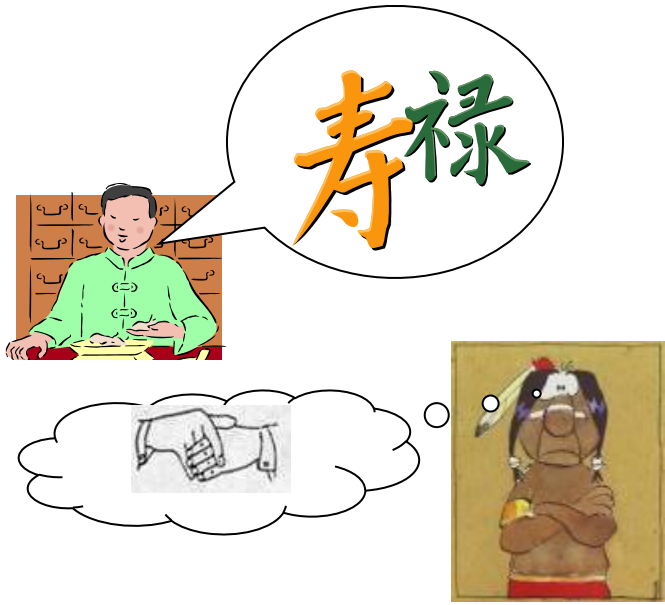


- Resolución de problemas con ciertas consultas
  - Ambigüedad
  - Variación lingüística
  - Contexto de la consulta
- Problemas
  - Usabilidad
  - Lenguaje natural y codificación semántica
  - Diferentes codificaciones semánticas

Web Semántica
Intercambio no ambiguo de información

- Para intercambiar información se necesita Interoperabilidad, compartiendo:
  - Un mismo medio con la misma forma de comunicación
    → Textos con lenguaje de marcado en el **Web**, p.e. HTML
  - Compartir un **Vocabulario**
    → Vocabularios de Metadatos
  - Compartir una Sintaxis y una misma forma de expresarse
    → p.e. XML/RDF

  Compartir un mismo conocimiento común
    → Tesauros y ontologías



Contexto: Web Semántica y RDF

La Web Semántica permite navegar a través de datos y semánticas. Requisitos:

1. Identidad
2. Accesibilidad (p.e. protocolos http)
3. Estructura (normalización, RDF)
4. Navegación (recursos enlazados)

El valor puede ser una URI o un literal

RDF está compuesto por tripletas con la forma:
SUJETO- ATRIBUTO- VALOR

Tripleta	Sujeto	Atributo	Valor
	www.john.com	Autor (URI)	John (URI/Literal)

Diferencia con XML

Posibilidades con XML:

```
<Libro idTitulo="El_Quijote"
xmlns="http://www.ejemplo.org/libro">
  <autor>Cervantes</autor>
</Libro>
```

```
<obra>
  <libro>El Quijote</libro>
<autor>Cervantes</autor>
</obra>
```

```
<autor nombre="Cervantes">
  <Libro>El Quijote</Libro>
</autor>
```

Pero con RDF…

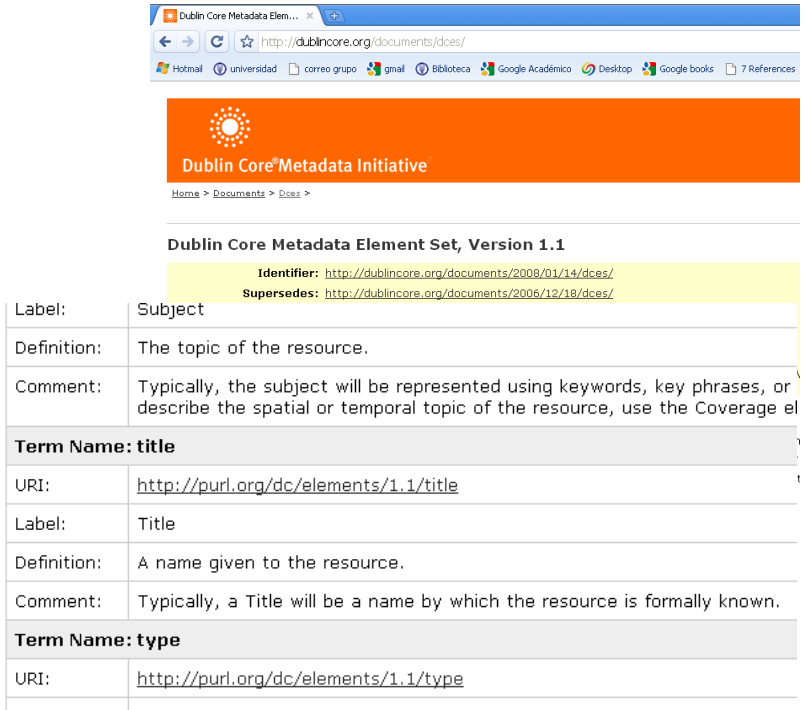
```
<rdf:RDF rdf:ID="El_Quijote"
        xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
        xmlns:dc="http://purl.org/dc/elements/1.1/"
         xml:base="http://www.mi_ejemplo.org/libro#">
        <rdf:Description ID="quijote">
                <dc:title>El Quijote</dc:title>
                <dc:creator>Miguel de Cervantes</dc:creator>
        </rdf:Description>
</rdf:RDF>
```

Namespaces
Ejemplo con Dublin Core y el elemento Title

```
<?xml version='1.0' encoding ='utf-8'
    standalone='yes'>
<Libros
    xmlns:dc="http://purl.org/dc/elements/1.1/title">
      <Libro>
          <dc:title lang="esp">Don
          Quijote de la Mancha</title>
          <Autor>Miguel de
              Cervantes</Autor>
      </Libro>
      <Libro>
          <dc:title lang="es">La vida es
              sueño</title>
          <Autor>Calderón de la Barca
              </Autor>
      </Libro>
```

namespace

http://purl.org/dc/elements/1.1/



Qname o prefijos del namespaces

Web 2.0

- Web 1.0, no interacción, estática
- Web 2.0
  - Interactividad: Entornos sociales y cooperativos. Espacios de opinión
  - Publicación más simple y estándar. Plantillas, lenguajes, código abierto, …
  - Descentralizado y flexible
  - Difusión y sindicación mejorada

Web 2.0

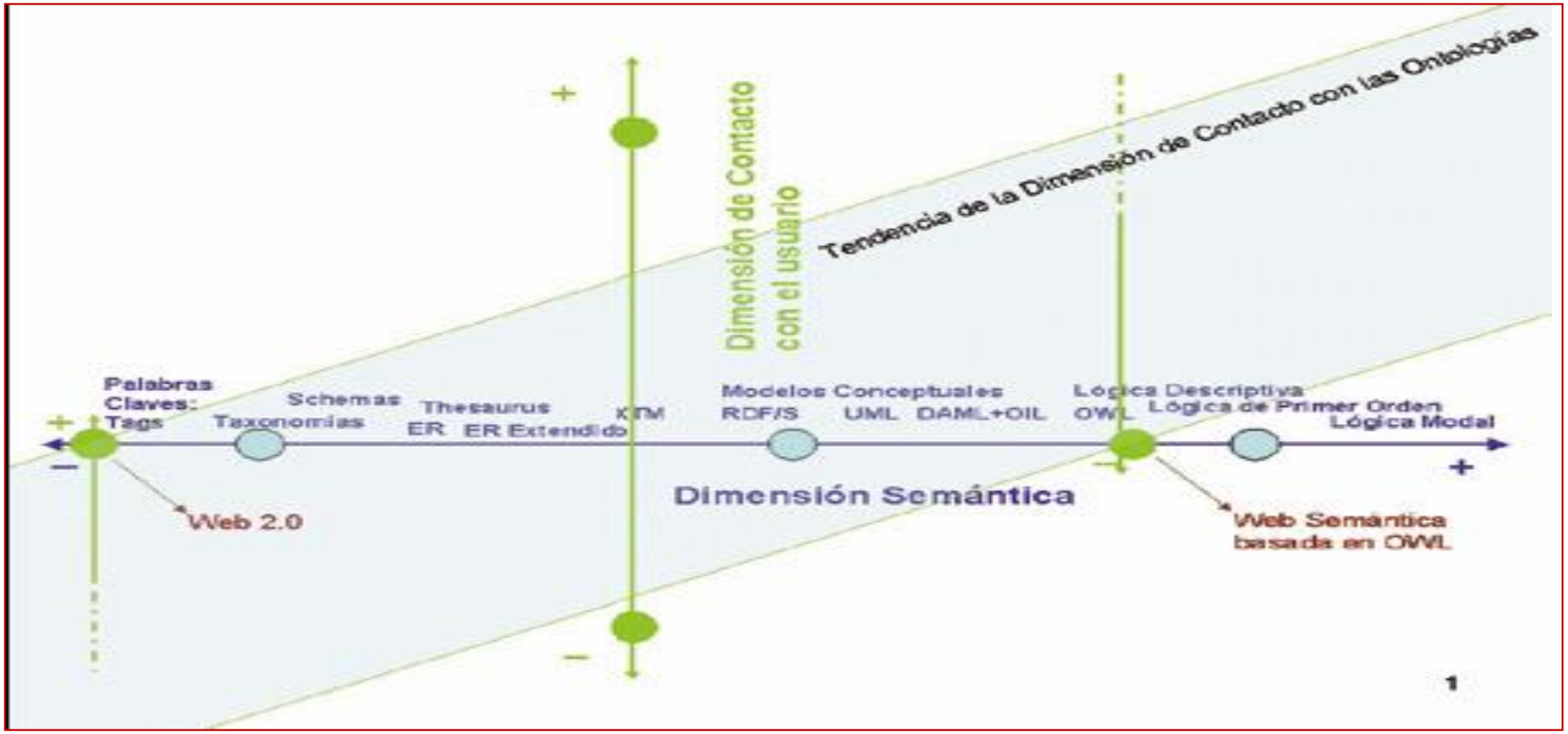
- Folksonomía descripción colaborativa y sin pautas preestablecidas
- Ejercicio indicar que problemas son previsibles (solución: ambigüedad, calidad, parcialidad)
- Descripciones para entornos sociales en la Web Social del.icio.us (enlaces favoritos), Flickr (fotos),Tagzania (lugares), o 43 Things
- Visualización: Wordle

Web 2.0 vs. Web Semántica

	Web 2.0	Web Semántica
Origen	Constatación de la evolución natural de la Web	Propuesta de Tim Berners para evolucionar la Web
Implantación	Muy alta	Escasa
Coordinación	No existe	Centralizada, sobre todo por el W3C
Foco	**Personas**	**Aplicaciones informáticas**
Creación	2003, 1ª conferencia 2004	1999 (Berners-Lee, 1999)

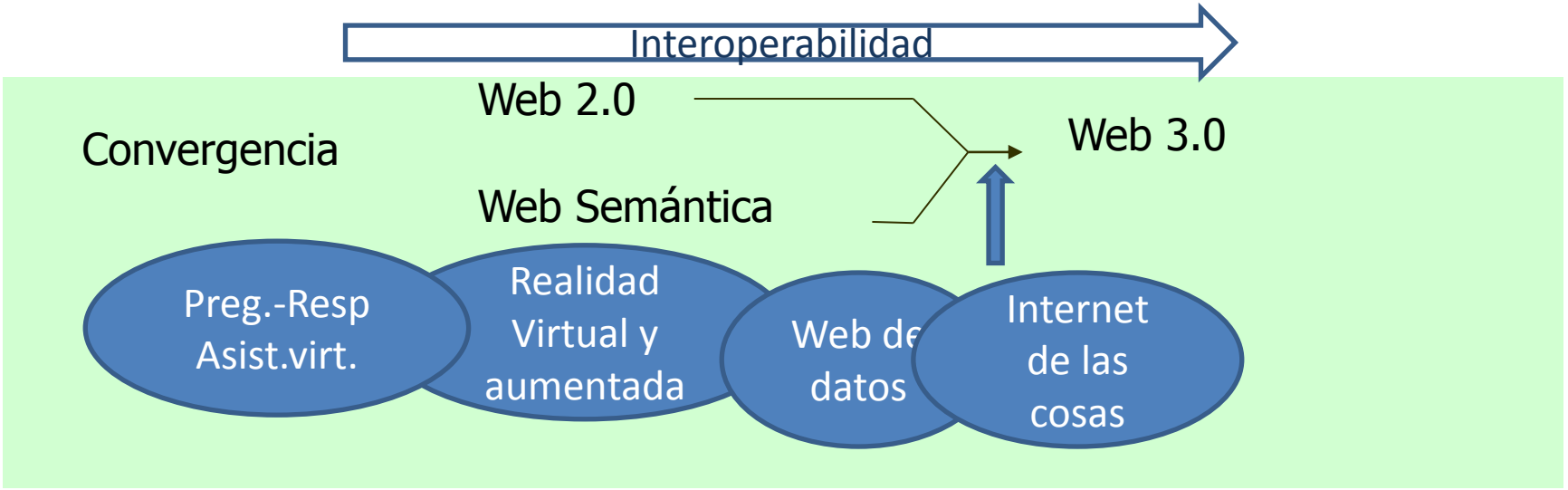
- Entran en confrontación por:
  - Las ontologías son poco legibles (RDF y OWL) por personas y costosas de crear. Las folksonomías son difíciles de interpretar por aplicaciones (polisemia y ambigüedad) pero su creación tiene bajo coste y esfuerzo
  - No hay herramientas de la Web Semántica amigables para los usuarios. Los recursos de la Web Social no son amigables para las aplicaciones
  - Técnicas automáticas de creación de ontologías inmaduras. Problemas de mapeo
  - Duplicidades de vocabularios de metadatos (p.e. SKOS-Core, los PSI, Zthes y MADS)

Web 2.0 vs. Web Semántica



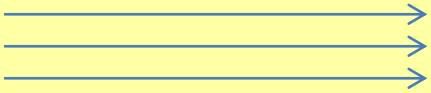
- La Web 3.0 intenta eliminar las dificultades de comunicación con la Web Semántica.
- La web de datos da mayor conocimiento a las aplicaciones web.
- La página web pasa a tener un papel menor, los datos se pueden usar y visualizar de distintas formas.
- El usuario puede leer, publicar pero también ejecutar nuevas aplicaciones.

Posibilidades evolución Web



La Web 3.0 es denominada por algunos Web Semántica, con la necesaria mejora en usabilidad. Los Web sites perderán protagonismo

Web 1.0
Web 2.0
Web Semántica



Siempre debe utilizarse la Web que provee la solución más económica con funcionalidad suficiente

¿Pero y la credibilidad y fiabilidad de los datos?

Universidad
Carlos III de Madrid

Módulo I
Fundamentos Recuperación en Internet

Colaboradores

J.Morato, V.Palacios

M.Marrero, S.Sánchez-Cuadrado, J.Urbano